This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

*Department of Biochemistry, University of Virginia, Charlottesville, VA 22908; and †Mathematical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gerald M. Rubin, December 2, 1987 (received for review September 17, 1987)

ABSTRACT **We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive derivative of the FASTP program, which can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The RDF2 program can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be displayed as a "graphic matrix" plot or as individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices.**

We have been developing tools for the analysis of protein and DNA sequence similarity that achieve a balance of sensitivity and selectivity on the one hand and speed and memory requirements on the other. Three years ago, we described the FASTP program for searching amino acid sequence data bases (1), which uses a rapid technique for finding identities shared between two sequences and exploits the biological constraints on molecular evolution. FASTP has decreased the time required to search the National Biomedical Research Foundation (NBRF) protein sequence data base by more than two orders of magnitude and has been used by many investigators to find biologically significant similarities to newly sequenced proteins. There is a trade-off between sensitivity and selectivity in biological sequence comparison: methods that can detect more distantly related sequences (increased sensitivity) frequently increase the similarity scores of unrelated sequences (decreased selectivity). In this paper we describe a new version of FASTP, FASTA, which uses an improved algorithm that increases sensitivity with a small loss of selectivity and a negligible decrease in speed. We have also developed a related program, LFASTA, for local similarity analyses of DNA or amino acid sequences. These programs run on commonly available microcomputers as well as on larger machines.

METHODS

The search algorithm we have developed proceeds through four steps in determining a score for pair-wise similarity. FASTP and FASTA achieve much of their speed and selectivity in the first step, by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison (2). The *ktup* parameter determines how many consecutive identities are required in a match. For example, if *ktup* = 4 for a DNA sequence comparison, only those identities that occur in a run of four consecutive matches are examined. In the first step, the 10 best diagonal regions are found using a simple formula based on the number of *ktup* matches and the distance between the matches without considering shorter runs of identities, conservative replacements, insertions, or deletions (1, 3).

In the second step of the comparison, we rescore these 10 regions using a scoring matrix that allows conservative replacements and runs of identities shorter than *ktup* to contribute to the similarity score. For protein sequences, this score is usually calculated using the PAM250 matrix (4), although scoring matrices based on the minimum number of base changes required for a replacement or on an alternative measure of similarity can also be used with FASTA. For each of these best diagonal regions, a subregion with maximal score is identified. We will refer to this region as the "initial region"; the best initial regions from Fig. 1*A* are shown in Fig. 1*B*.

The FASTP program uses the single best scoring initial region to characterize pair-wise similarity; the initial scores are used to rank the library sequences. FASTA goes one step further during a library search; it checks to see whether several initial regions may be joined together. Given the locations of the initial regions, their respective scores, and a "joining" penalty (analogous to a gap penalty), FASTA calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. FASTA uses the resulting score to rank the library sequences. We limit the degradation of selectivity by including in the optimization step only those initial regions whose scores are above a threshold. This process can be seen by comparing Fig. 1*B* with Fig. 1*C*. Fig. 1*B* shows the 10 highest scoring initial regions after rescoring with the PAM250 matrix; the best initial region reported by FASTP is marked with an asterisk. Fig. 1*C* shows an optimal subset of initial regions that can be joined to form a single alignment.

In the fourth step of the comparison, the highest scoring library sequences are aligned using a modification of the optimization method described by Needleman and Wunsch (5) and Smith and Waterman (6). This final comparison considers all possible alignments of the query and library sequence that fall within a band centered around the highest scoring initial region (Fig. 1*D*). With the FASTP program, optimization frequently improved the similarity scores of related sequences by factors of 2 or 3. Because FASTA calculates an initial similarity score based on an optimization of initial regions during the library search, the initial score is

---

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: NBRF, National Biomedical Research Foundation.

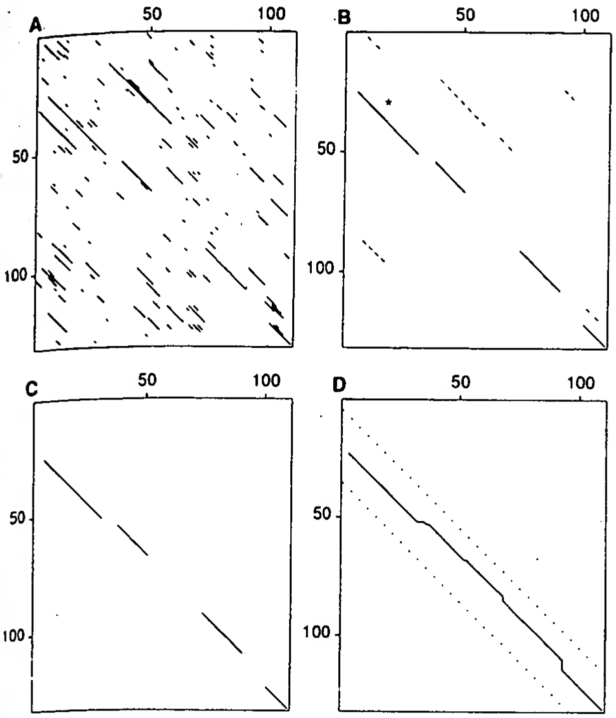


FIG. 1. Identification of sequence similarities by FASTA. The four steps used by the FASTA program to calculate the initial and optimal similarity scores between two sequences are shown. (*A*) Identify regions of identity. (*B*) Scan the regions using a scoring matrix and save the best initial regions. Initial regions with scores less than the joining threshold (27) are dashed. The asterisk denotes the highest scoring region reported by FASTP. (*C*) Optimally join initial regions with scores greater than a threshold. The solid lines denote regions that are joined to make up the optimized initial score. (*D*) Recalculate an optimized alignment centered around the highest scoring initial region. The dotted lines denote the bounds of the optimized alignment. The result of this alignment is reported as the optimized score.

much closer to the optimized score for many sequences. In fact, unlike FASTP, the FASTA method may yield initial scores that are higher than the corresponding optimized scores.

Local Similarity Analyses. Molecular biologists are often interested in the detection of similar subsequences within longer sequences. In contrast to FASTP and FASTA, which report only the one highest scoring alignment between two sequences, local sequence comparison tools can identify multiple alignments between smaller portions of two sequences. Local similarity searches can clearly show the results of gene duplications (see Fig. 2) or repeated structural features (see Fig. 3) and are frequently displayed using a "graphic matrix" plot (7), which allows one to detect regions of local similarity by eye. Optimal algorithms for sensitive local sequence comparison (6, 8, 9) can have tremendous computational requirements in time and memory, which make them impractical on microcomputers and, when comparing longer sequences, on larger machines as well.

The program for detecting local similarities, LFASTA, uses the same first two steps for finding initial regions that FASTA uses. However, instead of saving 10 initial regions, LFASTA saves all diagonal regions with similarity scores greater than a threshold. LFASTA and FASTA also differ in the construction of optimized alignments. Instead of focusing on a single region, LFASTA computes a local alignment for each initial region. Thus LFASTA considers all of the initial regions shown in Fig. 1*B*, instead of just the diagonal shown in Fig. 1*D*. Furthermore, LFASTA considers not only the band around each initial region but also potential sequence alignments for some distance before and after the initial region. Starting at the end of the initial region, an optimization (6) proceeds in the reverse direction until all possible alignment scores have gone to zero. The location of the maximal local similarity score in the reverse direction is then used to start a second optimization that proceeds in the forward direction. An optimal path starting from the forward maximum is then displayed (5). The local homologies can be displayed as sequence alignments (see Fig. 2*B*) or on a two-dimensional graphic matrix style plot (see Figs. 2*A* and 3).

Statistical Significance. The rapid sequence comparison algorithms we have developed also provide additional tools for evaluating the statistical significance of an alignment. There are approximately 5000 protein sequences, with 1.1 million amino acid residues, in the NBRF protein sequence library, and any computer program that searches the library by calculating a similarity score for each sequence in the library will find a highest scoring sequence, regardless of whether the alignment between the query and library sequence is biologically meaningful or not. Accompanying the previous version of FASTP was a program for the evaluation of statistical significance, RDF, which compares one sequence with randomly permuted versions of the potentially related sequence.

We have written a new version of RDF (RDF2) that has several improvements. (*i*) RDF2 calculates three scores for each shuffled sequence: one from the best single initial region (as found by FASTP), a second from the joined initial regions (used by FASTA), and a third from the optimized diagonal. (*ii*) RDF2 can be used to evaluate amino acid or DNA sequences and allows the user to specify the scoring matrix to be employed. Thus sequences found using the PAM250 scoring matrix can be evaluated using the identity or genetic code matrix. (*iii*) The user may specify either a global or local shuffle routine.

Locally biased amino acid or nucleotide composition is perhaps the most common reason for high similarity scores of dubious biological significance (10). High scoring alignments between query and library sequences may be due to patches of hydrophobic or charged amino acid residues or to A + T- or G + C-rich regions in DNA. A simple Monte Carlo shuffle analysis that constructs random sequences by taking each residue in one sequence and placing it randomly along the length of the new sequence will break up these patches of biased composition. As a result, the scores of the shuffled sequences may be much lower than those of the unshuffled sequence, and the sequences will appear to be related. Alternatively, shuffled sequences can be constructed by permuting small blocks of 10 or 20 residues so that, while the order of the sequence is destroyed, the local composition is not. By shuffling the residues within short blocks along the sequence, patches of G + C- or A + T-rich regions in DNA, for example, are undisturbed. Evaluating significance with a local shuffle is more stringent than the global approach, and there may be some circumstances in which both should be used in conjunction. Whereas two proteins that share a common evolutionary ancestor may have clearly significant similarity scores using either shuffling strategy, proteins related because of secondary structure or hydropathic profile may have similarity scores whose significance decreases dramatically when the results of global and local shuffling are compared.

Implementation. The FASTA/LFASTA package of sequence analysis tools is written in the C programming language and has been implemented under the Unix, VAX/VMS, and IBM PC DOS operating systems. Versions of the program that run on the IBM PC are limited to query se-

Table 1. FASTA and FASTP initial scores of the T-cell receptor (RWMSAV) versus the NBRF data base

NBRF code	Sequence	Initial score FASTA	Initial score FASTP
RWHUAV	T-cell receptor α chain	155	98
K1HURE	Ig κ chain V-I region	127	111
KVMS50	Ig κ chain V region	149	62
KVMSM6	Ig κ chain precursor V regions	141	64
KVRB29	Ig κ chain V region	126	54
L3HUSH	Ig λ chain V-III region	90	47
KVMS41	Ig κ chain precursor V region	87	87
RWMSBV	T-cell receptor β-chain precursor	94	94
RWHUVY	T-cell receptor β-chain precursor	91	59
RWHUGV	T-cell receptor γ-chain precursor	87	61
RWHUT4	T-cell surface glycoprotein T4	86	63
RWMSVB	T-cell receptor γ-chain precursor	71	41
HVMS44	Ig heavy-chain V region	67	36
G1HUDW	Ig heavy-chain V-II region	62	35

The average FASTP score = 26.1 ± 6.8 (mean ± SD). The average FASTA score = 26.2 ± 7.2 (mean ± SD). The mean and SD were computed excluding scores >54. V, Variable.

quences of 2000 residues; library sequences can be any length. Copies of the program are available from the authors.

Although FASTA and LFASTA were designed for protein and DNA sequence comparison, they use a general method that can be applied to any alphabet with arbitrary match/mismatch scoring values. All the scoring parameters, including match/mismatch values, values for the first residue in a gap and subsequent residues in the gap, and other parameters that control the number of sequences to be saved and the histogram intervals, can be specified without changing the program.

EXAMPLES

Comparison of FASTA with FASTP. To demonstrate the superiority of the FASTA method for computing the initial score, we compared the protein sequence of a T-cell receptor α chain (NBRF code RWMSAV) with all sequences in the NBRF protein data base[‡] and computed initial scores with both the present and previous methods. The T-cell receptor is a member of the immunoglobulin superfamily; in Release 12.0 of the data base, this superfamily has 203 members. FASTP placed 160 immunoglobulin superfamily sequences in the 200 top-scoring sequences; 57 related sequences received initial scores less than four standard deviations above the mean score. FASTA placed 180 superfamily members in the 200 top-scoring sequences; only 20 related sequences scored below four standard deviations above the mean. Table 1 contains specific examples from this data base search. Although there is often little difference in the two methods, this example shows that in a number of cases the new method obtains significantly higher scores between related sequences.

Nucleic Acid Data Base Search. FASTA can also be used to search DNA sequence data bases, either by comparing a DNA query sequence to the DNA library or by comparing an amino acid query sequence to the DNA library by translating each library DNA sequence in all six possible reading frames. We compared the 660-nucleotide rat transforming growth factor type α mRNA (GenBank locus RATTGFA) with all the mammalian sequences in Release 48 of GenBank[§]. We set *ktup* = 4 (see *Methods*), and the search was completed in under 15 min on an IBM PCAT microcomputer. The 10 top-scoring library sequences are shown in Table 2. Although it can be seen that the 3 top-scoring sequences are clearly related to RATTGFA, there are other high-scoring sequences that are probably not related, and the mouse epidermal growth factor, found in the translated data base search (Table 3), is not found among the top-scoring sequences.

Table 2. DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Score Initial	Score Optimized
HUMTFGAM	Human TGF mRNA	1336	1618
HUMTGFA2	Human TGF gene (exon 2)	354	366
HUMTGFA1	Human TGF gene (5′ end)	224	381
MUSRGEB3	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSRGE52	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSMHDD	MHC class I H-2D	122	78
HUMMETIF1	Metallothionein (MT)I$_F$ gene	116	92
MUSRGLP	45S rRNA (5′ end)	115	83
HUMPS2	pS2 mRNA	105	106
MUSC1A11	α-1 type I procollagen	86	89

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; MHC, major histocompatibility complex.

To further examine the similarity detected between RATTGFA and MUSRGEB3, a mouse rRNA gene cluster, we used the RDF2 program for Monte Carlo analysis of statistical significance (the window for local shuffling was set to 10 bases). Of the 50 shuffled comparisons (data not shown), 1 obtained an initial score greater than 140 (the observed initial score), and 9 shuffled sequences obtained optimized scores greater than 107 (the observed optimized score). Therefore, the similarity between RATTGFA and MUSRGEB3 is unlikely to be significant.

Translated Nucleic Acid Data Base Search. When searching for sequences that encode proteins, amino acid sequence comparisons are substantially more sensitive than DNA sequence comparisons because one can use scoring matrices like the PAM250 matrix that discriminate between conservative and nonconservative substitutions. A variant of FASTA, TFASTA, can be used to compare a protein sequence to a DNA sequence library; it translates the DNA sequences into each of six possible reading frames "on-the-fly." TFASTA translates the DNA sequences from beginning to end; it includes both intron and exon sequences in the translated protein sequence; termination codons are translated into unknown (X) amino acids. Table 3 shows the results of a translating search of the mammalian sequences in the GenBank DNA data base using the RATTGFA protein sequence as the query and *ktup* = 1. In the translated search, the mouse epidermal growth factor now obtains an initial score higher than any unrelated sequences; however, HUMTGFA1, which was found in the DNA data base search but only contains 13 translated codons, is no longer among the top scoring sequences.

Local Similarities. Fig. 2 displays the output of a local similarity analysis (*ktup* = 4) of CHPHBA1M, a chimpanzee α1-globin mRNA, and RABHBAP1, a rabbit α-globin gene, including the complete coding sequence and a flanking pseudo-θ_1-globin gene. LFASTA can either display a graphic matrix style plot of the local homologies (Fig. 2*A*) or the alignments themselves (Fig. 2*B*). The right-most three alignments (Fig. 2*A*) match the corresponding regions of the mRNA to exon subsequences from the pseudogene. We note that the FASTA initial score for the comparison of CHPH-

---

[‡] Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 12.

[§] EMBL/GenBank Genetic Sequence Database (1987) (Intelligenetics, Mountain View, CA), Tape Release 48.

Table 3. Translated DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Frame	Score	
			Initial	Optimized
RATTGFA	Rat TGF type α	1	816	816
HUMTGFAM	Human TGF mRNA	2	671	770
HUMTGFA2	Human TGF gene	1	204	205
MUSEGF	Mouse EGF mRNA	3	93	129
MUSMHAB3	Mouse MHC class II H2-IA$_\beta$	1	91	58
MUSIGCD17	Mouse Ig germ-line DJC region	3′	85	48
HUMESTR	Human estrogen receptor	3	83	65
RATINSI	Rat insulin 1 (*Ins-1*) gene	2	81	63
MUSTHYS1	Mouse thymidylate synthase	2	80	63
HUMPNU3	Human purine nucleoside phosphorylase	1′	80	52

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; EGF, epidermal growth factor; D, diversity; J, joining; C, constant; MHC, major histocompatibility complex.

BA1M and RABHBAPT would be based on the three globin gene exons, while the FASTP initial score would be based on a single conserved exon.

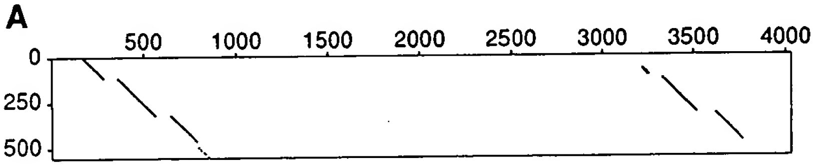
The Smith–Waterman optimization used in the LFASTA program allows the detection of more subtle features than can be detected by the eye using a graphic matrix plot, because the path traced is locally optimal, even though it may only have a slightly higher density of identities and conservative replacements. Fig. 3 shows a plot from a local similarity self-comparison of the myosin heavy chain from the nematode *Caenorhabditis elegans* (MWKW) using the PAM250 matrix. The amino-terminal half of the molecule forms a large globular head without any periodic structure; the solid line down the main diagonal represents the expected identity of the sequence with itself. The symmetrical parallel lines along the carboxyl-terminal half of the molecule correspond to the 28-residue repeat responsible for the α-helical coiled-coil structure of the rod segment.

DISCUSSION

In searching a data base, one is attempting to measure relatedness; in aligning two homologous sequences, one is trying to choose the most likely set of mutations since their divergence from a common ancestral sequence. Thus any tool for the analysis of sequence similarities must contain within it an implicit model of molecular evolution. An algorithm that guarantees the optimality of its alignments based on a set of scoring rules must be judged on how well these rules fit our current understanding of the process of molecular evolution. Algorithms that sacrifice realism to achieve greater efficiency, regardless of their mathematical rigor, require careful empirical evaluation.

Even though the tools we have developed use rigorous algorithms at each step and incorporate a realistic model of evolution, their hierarchical nature make them heuristic. The original FASTP program has had the benefit of extensive use and evaluation by a wide variety of scientists. The FASTA program exploits refinements of the previous approach that result in a significant improvement in sensitivity. The LFASTA local similarity analysis program is also a logical extension of the FASTP approach.

Because of the trade-offs between sensitivity and selectivity in data base searches, the results of any search, and particularly those that result in alignment scores that are not clearly separated from the distribution of all library sequence



B

```
               10        20        30        40        50        60
CHPHBA GACTCAGAAAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG
       :::: :::: ::: X:::::::::::::::::: :: :::::::::::: ::::: : :
RABHBA GACTGAGAAGGAA-CCACCATGGTGCTGTCTCCCGCTGACAAGACCAACATCAAGACTG
              180        190       200       210       220

               70        80        90       100       110
CHPHBA CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGG
       ::::::  ::: :::::   :::: :::::::::::: :: :::::: :::::::X
RABHBA CCTGGGAAAAGATCGGCAGCCACGGTGGCGAGTATGGCGCCGAGGCCGTGGAGAGG
      230       240       250       260       270       280
```

FIG. 2. Local comparison of an α-globin mRNA sequence with an α-globin gene cluster. An ape α_1-globin mRNA sequence (GenBank sequence CHPHBA1M) was compared with a rabbit α-globin gene sequence (RABHBAPT) containing a second pseudo-θ-globin gene using the LFASTA program. (*A*) A plot of the homologous regions shared by the two sequences. (*B*) One of the alignments between the mRNA sequence and the rabbit α-globin gene (nucleotides 171–855). Three other alignments between the mRNA sequence and the α-globin gene and three alignments between the pseudo-θ-globin gene (nucleotides 3200–3770) were calculated but are not shown. There is 84.3% identity in the 115 nucleotide overlap. The initial region and optimized scores using LFASTA are 284 and 304, respectively. X denotes the ends of the initial region found by LFASTA.

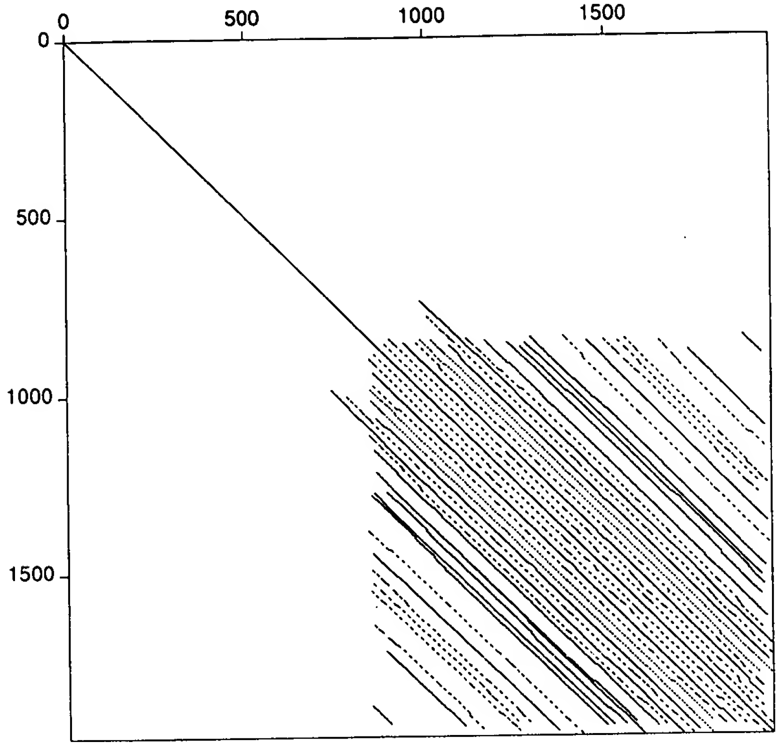


FIG. 3. Repeated structure in the myosin heavy chain. LFASTA was used to compare the *Caenorhabditis elegans* myosin heavy chain protein sequence (NBRF code MWKW) with itself using the PAM250 scoring matrix. The solid, dashed, and dotted lines denote decreasing similarity scores. The solid lines had initial region scores greater than 80 and optimized local scores greater than 150; the longer dashed lines had initial region and optimized local scores greater than 65 and 120, respectively, and the shorter dashed lines had initial region and optimized local scores greater than 50 and 100, respectively. Homologous regions with lower scores are plotted with dots.

scores, must be carefully evaluated (1, 11). The Monte Carlo analysis of statistical significance provided by a program such as RDF2 can often be critical in evaluating a borderline similarity. Previously we suggested ranges of z values [(observed score − mean of shuffled scores)/standard deviation of shuffled scores] corresponding to approximate significance levels. However the z values determined in a Monte Carlo analysis become less useful as the distribution of shuffled scores diverges from a normal distribution, as is found with FASTA. Therefore, we now focus on the highest scores of the shuffled sequences. For example, if in 50 shuffled comparisons, several random scores are as high or higher than the observed score, then the observed similarity is not a particularly unlikely event. One can have more confidence if in 200 shuffled comparisons, no random score approaches the observed score. In general, our experience has led us to be conservative in evaluating an observed similarity in an unlikely biological context.

These programs provide a group of sequence analysis tools that use a consistent measure for scoring similarity and constructing alignments. FASTA, RDF2, and LFASTA all use the same scoring matrices and similar alignment algorithms, so that potentially related library sequences discovered after the search of a sequence data base can be evaluated further from a variety of perspectives. In addition, LFASTA can also show alternative alignments between sequences with periodic structures or duplications.

1. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
2. Dumas, J. P. & Ninio, J. (1982) *Nucleic Acids Res.* **10**, 197–206.
3. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
4. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
5. Needleman, S. & Wunsch, C. (1970) *J. Mol. Biol.* **48**, 444–453.
6. Smith, T. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
7. Maizel, J. & Lenk, R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665–7669.
8. Goad, W. & Kanehisa, M. (1982) *Nucleic Acids Res.* **10**, 247–263.
9. Sellers, P. H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3041.
10. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
11. Doolittle, R. (1981) *Science* **214**, 149–159.